What is a Valid Scientific Study?
An Analysis of Selection Criteria Used by the What Works Clearinghouse

Technical Report 2014-3

Jean Stockard, Director of Research and Evaluation, NIFDI

August 25, 2014

Table of Contents

List of Tables

Executive Summary[1]

The educational research literature includes thousands of studies regarding the efficacy of educational programs. Numerous meta-analyses and reviews have summarized this material. The literature on Direct Instruction programs is especially large. For instance, in a review of meta-analyses, Hattie (2009) reported the results of 304 studies of Direct Instruction, involving over 42,000 students. Despite this very large literature base, the What Works Clearinghouse has identified very few studies that meet their selection criteria and evidence standards. For instance, in a November 2013 review of *Reading Mastery (RM)* for Beginning Readers, they reported finding no studies published since 1983 that could meet their standards. Why, given such a large literature base, does the WWC accept so few studies for review?

This paper, and a companion one (NIFDI Technical Report 2014-4), examine that question. The results indicate that the differences involve two general issues: 1) the policies of the WWC and the ways in which they differ from those typically used within the social sciences and 2) errors in the review procedures of the WWC. The companion report examines specific errors in the recent WWC analysis of *RM* and describes the conclusions that would have been made if these errors had not occurred and if standard methodological criteria had been used. This report focuses on the policies of the WWC, contrasting their procedures with traditional and standard methodological approaches.

The first part of the report compares the WWC's procedures with those that are standard within the social sciences. It contrasts the WWC's exclusive approach to the much more inclusive approach commonly used in the social sciences and typified by the writings of Campbell and Stanley and their successors. For instance, the WWC accepts only studies that use a pretest-posttest control group design, giving the highest ranking only to those that include random assignment. In contrast, the standard methodological tradition recommends a much wider range of designs, especially in organizational settings such as schools. This standard literature offers a number of alternatives to random assignment, noting that random assignment is often not feasible and, in fact, can diminish the internal validity of a study. The standard tradition also downplays the need for pretesting, showing a number of viable alternatives. In short, while the WWC has a very limited definition of acceptable research designs, the standard methodological literature provides a flexible and broad list of approaches that can and should be used in field settings such as those in education.

[1] The author gratefully acknowledges the extraordinarily skilled assistance of Timothy Wood in compiling information for this report and the helpful comments of Muriel Berkeley, Douglas Carnine, Christina Cox, Gary Davis, Siegfried Engelmann, and Jerry Silbert on earlier drafts. All conclusions and opinions in this document are, however, the sole responsibility of the author.

Even if a study uses a pretest-posttest control group design, it must pass additional standards to be considered by the WWC. Two of these, discussed in the first section of this report, appear to be especially difficult to meet. The first involves the requirement that pretest differences between the intervention and control condition fall within .25 of a standard deviation of each other on all measures. The randomized trials preferred by the WWC are often small and the WWC requires, for full endorsement, that numerous measures be included. Yet, simply by chance, the probability of meeting the WWC standard for group equivalence declines as samples become smaller and as more measures are included. The second, often problematic, standard involves the "one unit" rule, which requires that studies include data from more than one classroom, school, or district. This rule applies even if a classroom has multiple teachers, a school has multiple classrooms or data from multiple years, or a district has multiple schools. A large proportion of studies are automatically discarded based on this standard. Interestingly, the WWC provides no justification for these standards from the research literature.[2]

The result of this process is that the WWC finds very few, and often no, studies that are deemed worthy of consideration for a given report. A thought experiment illustrates why it is so difficult to meet the WWC's standards. It postulates characteristics of a "perfect" study that would include the preferred pretest-posttest control group with randomized assignment design; incorporate sufficient districts, schools, and students to pass the "one unit" standard; and have enough subjects for a reasonable probability of meeting the standards of group equivalence. Such a study could be extraordinarily expensive and would have to be quite large. Moreover, given statistical realities, it would have no guarantee of passing the WWC standards involving group equivalence. While each WWC criterion and standard might on its own appear logical, collectively they result in such severe restrictions that the probability of any study meeting all of these standards is very remote.

The standard methodological literature states quite firmly that there can never be a "perfect" experiment and instead calls for numerous tests of hypotheses, involving different settings, samples, and designs. This is the classic notion of a cumulative science. The second section of this report illustrates the use of such an accumulation of findings by examining results of 37 studies of the *Reading Mastery* curricular program with primary aged students. All of these studies were identified in the November 2013 WWC report but were excluded from consideration for one or more of the reasons described above. Results across these studies were very consistent. The average effect was more than twice the standard generally used to denote educational importance. This finding replicates the research literature on *RM*.

[2] As described in the text, the except was a citation given to justify the group equivalence criterion, but it actually addressed a different issue and did not justify the criterion.

Multivariate statistical analyses were then used to examine the extent to which application of the WWC criteria affects substantive conclusions regarding the program's efficacy. The results indicate that there are no differences in the conclusions. There is no indication that effects are systematically smaller or larger when studies meet or do not meet a given WWC criterion. The effects associated with *Reading Mastery* remained statistically significant and substantively strong when any of the WWC exclusion criteria were used. There seems to be no value added to the estimates of *RM*'s efficacy by considering the factors deemed important by the WWC.

On the other hand, the costs associated with the restrictions used by the WWC are far from minimal. As implied above, developing a study that would have a modest probability of passing the WWC standards could be extraordinarily expensive. In addition, the costs associated with the WWC screening process are no doubt very high, with what appear to be a large number of person hours devoted to assessing studies for their eligibility. If the criteria were altered to more closely mirror those that are standard in the methodological literature, these resources could be reallocated to the review and summation of research findings as well as to more extensive checking of results.[3]

Just as important are the costs to the public. The use of the WWC's procedures deprives the public of a true view of the state of educational research. By suggesting that there are very few studies worth considering, they dismiss decades of solid research, much of it funded by the federal government and published in well regarded journals in the field. Moreover, by devoting attention to only a very few studies out of the many that are available, the WWC provides only a very small glimpse of the body of research results. Simple sampling theory would suggest that this extraordinarily small sample could easily provide biased results, further justification for the use of the more inclusive approach described within the classic and standard methodological literature.

Given that the WWC's selection procedures do not appear to be justified by scholarly research or by empirical evidence it seems reasonable to suggest that they should be reconsidered. The WWC was funded to provide a cumulative picture of what works in education. But, the selection criteria that it employs embody such severe restrictions that it is virtually impossible to build a cumulative understanding of results. The standard methodological literature has established and reliable techniques and procedures for developing these cumulative understandings. The educational community, students and their parents would be well served if the WWC would adopt these standard procedures.

[3] The companion technical report (2014-4) documents a large number of errors at all stages of the review process in the November 2013 on RM, suggesting that much more attention needs to be paid to issues of accuracy in the review process. A forthcoming NIFDI Technical Report (2014-5) documents that numerous other organizations and groups have found errors and objected to WWC policies.

What is a Valid Scientific Study?
An Analysis of Selection Criteria
Used by the What Works Clearinghouse

The What Works Clearinghouse (WWC) is a federally funded program established in 2002 to evaluate educational interventions and provide summary ratings and reports of their efficacy. The WWC's website describes their organization as a "trusted source of scientific evidence for what works in education to improve student outcomes" and as providing "accurate information on education research" (WWC, 2013a). Yet, the reports of the WWC often directly contradict those within the research literature. While literature reviews and meta-analyses report on literally hundreds of studies regarding educational programs, the WWC reports that it can find very few studies that it considers worthy of consideration. Such contradictions are associated with several reports of Direct Instruction programs, including a report on *Reading Mastery* issued by the WWC in November, 2013 (WWC, 2013b).

Reading Mastery (RM) is a reading program that is part of the Direct Instruction corpus of curricula. A large body of literature has documented the effectiveness of Direct Instruction programs in promoting achievement. In a recent review Coughlin (2014) summarized several extensive summaries of these studies, noting the consistent and strong support for the programs' efficacy (e.g. Adams and Engelmann, 1996; Kinder, et al., 2005; Przychodzin-Havis, et al., 2005; Schieffer, et al., 2002). Authors of these analyses have commented on the extensive nature and the quality of this literature base, especially in comparison to other programs (e.g. Borman, et al., 2003, p. 141; Hattie, 2009, pp. 206-207). Yet, the WWC reported that it could find "no studies of *Reading Mastery* that fall within the scope of the Beginning Reading review protocol [and] meet What Works Clearinghouse (WWC) evidence standards" (WWC, 2013b, p. 1). Why did this discrepancy occur? Given the large literature about Reading Mastery, how could the WWC find no studies of its efficacy?

This Technical Report and a companion report (NIFDI Technical Report 2014-4) examine those questions. The results indicate that the discrepancies result from two general issues: 1) the policies of the WWC and the ways in which they differ from those typically used within the social sciences and 2) errors in the review procedures of the WWC. This report focuses on the first issue – the WWC's policies and procedures. The companion report examines specific errors in the recent WWC analysis of *RM* and describes the conclusions that would have been made if these errors had not occurred and if standard methodological criteria had been used.[4]

[4] A forthcoming NIFDI Technical Report (2014-5) provides extensive documentation of the issues that others have found with the WWC policies and procedures.

The first section of this report contrasts WWC policies and criteria to the standard literature on research design. It describes numerous ways in which the WWC practices deviate from standard methodological procedures and how application of the WWC policies results in a very limited view of the research literature. The second section of the report presents a quantitative analysis of the impact and utility of the WWC's criteria, analyzing the extent to which the application of their standards affects summaries of findings. In other words, does the use of the WWC criteria provide improved estimates of the effectiveness of a program? The analysis concludes that the WWC criteria provide no added value to the validity of effect estimates. Instead, they result in a very limited and, thus potentially biased, view of the body of literature. A final section discusses the implications of the analysis for both the policies and the procedures of the WWC. It suggests that altering the WWC policies to more closely align with those of the traditional methodological literature would provide a more accurate view of the research base and be more helpful to practitioners and policy makers.

WWC Policies and Traditional Approaches to Research Design

The WWC process for selecting studies to review involves three separate stages. First, a list of relevant literature is developed. Second, the identified materials are screened to determine if they are eligible for review within the protocol for a given review. Third, studies that meet the general eligibility criteria are examined to see if they meet WWC methodological standards of evidence. Only studies that pass this third step are then reviewed. As noted above, very few studies appear to get through this winnowing process.

This section examines some of the criteria involved in the second and third steps of the WWC process to try to understand why so few studies are selected for review. First, the WWC approach is situated within general methodological traditions regarding research design, contrasting the WWC's exclusive definition of valid research designs with the more inclusive definition usually used by social scientists. Then several specific WWC criteria used at both the second and third steps of the review process are discussed.[5] A summary section examines the cumulative impact of the various WWC criteria on the probability that a study would be selected for review.

Exclusive and Inclusive Definitions of Experimentation

Taken together, the WWC criteria and standards appear to involve an exclusive definition of appropriate research designs. They describe a "perfect" experiment that reflects, at least implicitly, the writings of the British statistician, R.A. Fisher (1925, 1935), one "in which an experimenter having complete mastery can schedule treatments and measurements for

[5] This is a subset of the standards and the criteria used by the WWC and reflect those that most often impacted the 2013 review of *RM*.

optimal statistical efficiency" (Campbell & Stanley, 1963, p. 1).[6] Such a highly controlled approach has been called the "gold standard" for its tight control of any and all factors apart from the experimental variable that could affect outcomes. In reality, of course, the social sciences, which deal with living, interacting people in groups and complex organizations, comprise an environment far removed from a tightly controlled, isolated laboratory setting.

It is not surprising then that the standard methodological approach to experimentation within the social sciences, typified by the writings of Campbell and Stanley (1963) and their successors (Cook & Campbell, 1979; Shadish, Cook, and Campbell, 2002), embodies a flexible approach to research design. Three elements of these works, referred to here as the Campbell, Cook, Shadish, and Stanley (CCSS) tradition, are especially relevant to this discussion.

First, the CCSS tradition describes a wide variety of research designs that could be appropriate for field settings, such as schools. It emphasizes the importance of internal validity, determining if in fact the experimental treatments make a difference. This is, of course, the aim of the Fisher tradition of tightly controlled experiments. But, the CCSS writings show how a wide variety of designs can be internally valid.

Second, the CCSS tradition also emphasizes external validity, the "question of generalizability: To what populations, settings, treatment variables and measurement variables can this effect be generalized" (Campbell & Stanley, 1963, p. 5). In this context they specifically discuss the importance of promoting both external and internal validity in research on schools and teaching:

> [T]he selection of designs strong in both types of validity is obviously our ideal. This is particularly the case for research on teaching, in which generalization to applied settings of known character is the desideratum (Campbell & Stanley, 1963, p. 5).

Third, the CCSS discussions, especially in the later volumes in the tradition, also focus on the cumulative nature of science and the importance of systematically contrasting the results of multiple tests across varying samples, settings, and outcome measures. As Cook and Campbell put it,

> we stress the need for *many* tests to determine whether a causal proposition has or has not withstood falsification; such determinations cannot be made on one or two failures to achieve predicted results (1979, p. 31, emphasis in original).

[6] The WWC writings do not cite either the CCSS tradition or Fisher and, surprisingly, have very few academic citations.

They propose a "grounded theory of generalized causal inference," which they suggest "is more practical than random sampling for daily scientific work," and builds on numerous conceptualizations developed over the last half century (Shadish, et al., 2002, p. 348, citing Brunswick, 1956, Campbell, 1986, Cook, 1990, 1991; Cronbach, 1982; Cronbach & Meehl, 1955). Like meta-analyses and systematic literature reviews, the grounded theory of generalized causal inferences summarizes findings across many studies, studies that involve a wide variety of samples, settings, outcome measures, and designs. Using the Popperian notion of falsification, hypotheses that are proven false are discarded, and those that receive support (or technically have not yet been falsified) are retained (Cohen, 1989; Popper, 1962). This systematic accumulation of findings and the careful examination of variations in results are crucial in determining external validity. This approach is used in the discussion below regarding the accumulation of findings regarding *Reading Mastery*. Next, however, the criteria used by the WWC are described and contrasted to those traditionally used by social scientists.

Acceptable Designs: Criteria at Stage Two of the WWC Process

As noted above, once the WWC has amassed a list of articles regarding an intervention, the studies are examined to see if they employ an "acceptable" research design. The WWC's exclusive approach to judgment of research is reflected in their stipulations regarding valid research designs. These stipulations contrast strongly with the traditional literature. Three general areas are important: the role of randomized assignment, the use of pretests, and the range of research designs appropriate for educational settings.

Randomized Assignment – The WWC description of acceptable designs is as follows:

> The WWC includes findings from studies of effectiveness that use a comparison group that was created randomly (randomized controlled trials) or through a process that was not random (quasi-experimental designs)….Studies using other study designs are not eligible for review" (WWC, 2014, pp. 7-8).
>
> Randomized controlled trials can receive the highest WWC rating of *Meets WWC Group Design Standards without Reservations*. The distinguishing characteristic of a randomized controlled trial is that study participants are assigned randomly to form two or more groups that are differentiated by whether or not they receive the intervention under study (WWC, 2014, p. 9, emphasis in original).
>
> Quasi-experimental design studies that demonstrate baseline equivalence can receive a WWC rating no higher than *Meets WWC Group Design Standards with Reservations*. A quasi-experimental design compares outcomes for students, classrooms, or schools who had access to the

> intervention with those who did not but were similar on observable characteristics (WWC, 2014, p. 10, emphasis in original).[7]

Thus, it appears the only studies that can be accepted "without reservations" by the WWC are those that employ a pretest-posttest control group design with random assignment, one of the three "true experimental designs" described by Campbell and Stanley, the first volume in the CCSS series (1963, pp. 13-27). The only studies that can be accepted by the WWC "with reservation" are those using a pretest-posttest control group design, as long as they meet criteria regarding group equivalence (discussed in the next section).

The WWC's emphasis on randomized control trials and its limited acceptance of quasi-experimental designs contrasts sharply with the CCSS tradition. The CCSS tradition describes the importance of randomized experiments in establishing "causal description" (Shadish, et al, 2002, p. 9). Yet, the authors are also quite clear in stressing the importance of using research designs other than randomized experiments, noting that "experiments are far from perfect means of investigating causes" (Shadish, et al, 2002, p. 8):

> Among scientists, belief in the experiment as the *only* means to settle disputes about causation is gone, though it is still the preferred method in many circumstances. Gone, too, is the belief that the power experimental methods often displayed in the laboratory would transfer easily to applications in field settings (Shadish, et al, 2002, p. 30, emphasis in original).

In support of this point, the authors of later books in the CCSS tradition rued the use of the term "true experiment" in the 1963 volume, stating, "We shall not use the term [true experiment] at all given its ambiguity and given that the modifier *true* seems to imply restricted claims to a single correct experimental method" (Shadish, et al, 2002, p. 13). The authors described their work as

> about improving the yield from experiments that take place in complex field settings, both the quality of causal inferences they yield and our ability to generalize these inferences to constructs and over variations in persons, settings, treatments, and outcomes (Shadish, et al, 2002, p. 32).

In other words, the CCSS tradition emphasizes the possibility of valid research designs appropriate for the full range of settings in which humans interact. The authors conclude that even relatively complex models such as random selection of units from a population followed by random assignment to treatments

> cannot be advocated as *the* model for generalized causal inference....Though we unambiguously advocate it when it is feasible, we obviously cannot rely on

[7] The WWC also may choose to describe results from studies that use a regression discontinuity design or single-case, but the standards for these analyses are apparently not yet fully developed (WWC, 2014, pp. 8-9).

> it as an all-purpose theory of generalized causal inference. So researchers must use other theories and tools to explore generalized causal inferences of this type (Shadish, et al, 2002, p. 348, emphasis in original).

Pretesting – The role of pretesting is also less prominent in the CCSS literature than in the WWC criteria. Interestingly, the CCSS definition of randomized experiments does not mention pretests, focusing instead on the assignment of units to conditions and posttest assessment:

> The basic randomized experiment requires at least two conditions, random assignment of units to conditions, and posttest assessment of units (Shadish, et al, p. 257).

In fact, four of the nine examples of randomized designs listed in a summary table in the latest CCSS volume do not include pretests (see Shadish, et al, 2002, p. 258).

As in other areas of their discussion of research design, the CCSS authors take a practical approach to this issue. While noting that pretesting is sometimes "seriously impractical," they also note that "pretreatment information, preferably on the same dependent variable used at posttest, helps enormously in answering [questions about attrition of subjects]" (Shadish, et al, 2002, p. 260). While issues of group equivalence certainly become more complex when random assignment is not used, the CCSS writings also provide a number of examples of quasi-experimental designs that incorporate control groups without a pretest of all subjects (see Shadish, et al, 2002, pp. 115-130). Numerous authors have built on this logic in discussing research in educational settings and have suggested that restricting analyses of educational interventions to randomized control trials or the pretest-posttest control group design, as advocated by the WWC, may not be appropriate and could result in misleading conclusions (Cook, Scriven, Coryn, & Evergreen, 2009; McMillan, 2007; Raudenbush, 2008; Scriven, n.d., 2008; Slavin, 2008; Stockard, 2013c).[8]

These misleading conclusions are simply a function of trying to conduct controlled experiments in real-life organizational settings. The classic "gold standard" experiment in a field such as medicine is "double blind," with neither the interventionists (e.g. doctors) nor the subjects (e.g. patients) knowing which group they are in. In contrast, such a situation is

[8] The analysis in this paper parallels a growing body of scholarship that suggests that there has been an over-emphasis on randomized control trials in educational research, as well as other field settings. One element of this literature is empirical in nature and has compared the results obtained with randomized trials and quasi-experimental designs. Some have found systematic differences between these types of studies (e.g. Agordini & Dynarski, 2004; Glazerman, Levy, & Myers, 2003; Weisburd, Lum, & Petrosino's 2001), while others suggest that the results are minimal when appropriate statistical controls are employed (Heinsman & Shadish, 1996; Shadish, Clark, & Steiner, 2008; Slavin, 2008). The reviews suggesting differences between the techniques examined studies on criminal justice, welfare and job training programs, and dropout preventions. Those suggesting more minimal effects examined issues related to student achievement and learning, arguably the substantive area that would be of more concern to educational researchers.

generally impossible to attain in a field setting such as a school or even a school district with multiple schools. When teachers and students know that they are part of an experiment this knowledge becomes another potential causal variable in the design. There is, logically, no way for an analyst to determine if differences between the control and experimental group are due to the intervention or to other types of actions by the subjects based on their knowledge of the situation. Methodologists call this the Hawthorne effect or the John Henry effect (Zdep & Irvine, 1970; see also Engelmann, 2014).

Alternatives to Randomized Control Trials – The WWC criteria regarding acceptable designs quoted above indicate that the only quasi-experimental design that can be accepted by the WWC, albeit "with reservation," is the pretest-posttest control group design" provided that the experimental and control group are equivalent (see discussion below on this criterion). Yet, the CCSS tradition discusses numerous alternatives, both to randomized control trials and to the pretest-posttest control group design. Several of these alternatives can be especially useful in educational research.

The first involves what CCSS refer to as "recurrent institutional cycle" or "cohort control group" designs. They describe how this design is a useful alternative to randomized control trials in organizational settings and, especially, schools:

> Many institutions experience regular turnover as one group "graduates" to another level and their place is taken by another group. Schools are an obvious example of this, as most children are promoted from one grade to the next each year....The term cohort designates the successive groups that go through processes such as these. Cohorts are particularly useful as control groups *if* (1) one cohort experiences a given treatment and earlier or later cohorts do not; (2) cohorts differ in only minor ways from their contiguous cohorts; (3) organizations insist that a treatment be given to everybody, thus precluding simultaneous controls and making possible only historical controls; and (4) an organization's archival records can be used for constructing and then comparing cohorts (Shadish, Cook, and Campbell, 2002, pp. 148-149, emphasis in original,; see also Cook & Campbell, 1979, pp. 126-127; Campbell & Stanley, 1963, pp. 56-61; Stockard, 2013c).

The companion report (NIFDI Technical Report 2014-4) gives several examples of the use of this design in studies of the efficacy of *Reading Mastery*.

The second alternative involves what CCSS term "normed comparison contrasts," in which

> "obtained performance of the treatment group at pretest and posttest is compared with whatever published norms are available that might shed light on a counterfactual inference of interest....This form of comparison is also routinely used in educational studies to assess whether a group of students,

> classrooms, or schools rises over time in its percentile ranking on some published test. The possible rankings are taken from published norms and are meant to reflect the students' performance and its change relative to how students in the original norming sample did (Shadish, et al., 2002, pp. 126-7).

Interestingly, the normed comparison design was, at one time, promoted by the U.S. Department of Education (Tallmadge, 1977), which funds the WWC. The CCSS authors note limitations of normed comparison designs, especially when the treated groups depart markedly from the overall mean, but data from such studies certainly provide additional information about interventions, such as that used in analyses of the accumulation of findings, and thus should not be automatically dismissed. Again, several examples of the use of this design are given in the companion report.[9]

A third alternative is interrupted time series, described by Shadish and associates as "one of the most effective and powerful of all quasi-experimental designs" (Shadish, et al., 2002, p. 171). Time series refers to observations that occur consecutively over time, and "interrupted time series" refers to the impact of a treatment (the interruption) on the trend in the series. Time series analyses can involve one unit, such as one school or district, or multiple units. Several authors have described how a multiple base-line interrupted time-series approach may be especially suited to long-term evaluations of large group interventions (Biglan, Ary, & Wagenaar, 2000; Biglan, Flay, Komro, Wagenaar, & Kjelistrand, 2012; Madigan & Cross, 2012). An example of such an analysis of *Reading Mastery,* not included in the WWC's November 2013 report, is found in Stockard (2013a).

WWC Standards of Evidence: Criteria at Stage Three of the Review Process

If studies are found to fit within a given WWC review protocol, including having an acceptable design, they are then reviewed to see if they meet the WWC "standards of evidence." Issues related to two general standards appear to have affected decisions in the Fall 2013 report on *RM*: criteria for determining group equivalence and concerns about potential confounding influences. Again, the WWC's criteria regarding these issues contrast with accepted methodological standards.

Group Equivalence – Valid comparisons between intervention and comparison groups require, of course, that the two groups be equivalent. To ensure that such equivalence occurs, the WWC uses stringent guidelines on the extent of allowable differences at pretest:

> If the reported difference of *any* (emphasis added) baseline characteristic is greater than .25 standard deviations in absolute value (based on the variation of that characteristic in the pooled sample), the intervention and comparison groups are judged to be not equivalent. The standard limiting pre-intervention

[9] In a well-designed analysis, Tallmadge (1982) found norm referenced designs yield estimates of student gains "that are reasonably comparable to those derived from the randomized control group design (Tallmadge, 1982, p. 110).

> differences between groups to 0.25 standard deviations is based on Ho, Imai, King, and Stuart (2007). For differences in baseline characteristics that are between .05 and .25 the analysis must include a *statistical adjustment* (emphasis in original) for the baseline characteristics to meet the baseline equivalence requirement. Differences of less than or equal to 0.05 require no statistical adjustment (WWC, 2014, p. 15).

In other words, if any baseline (pretest) characteristic, such as the average score, of an experimental and control group differ by more than .25 of a standard deviation, the study is excluded from consideration. If the difference falls between .05 and .25 of a standard deviation, statistical adjustments must be used. Other requirements to establish baseline equivalence (WWC, 2014, pp. 15-16) include the presence of either pre-intervention measures that are "analogous" to post-intervention measures or the use of control variables specified by the WWC and demonstrating equivalence separately for each outcome. This requirement applies to studies that have employed random assignment as well as those that have used other means of assigning subjects to groups (see examples in Appendix A).

A simple application of sampling theory and the Central Limit Theorem shows how difficult it is to meet this criterion. The probability of having group differences smaller than .25 of a standard deviation declines sharply when multiple measures or comparisons are used and/or when samples are smaller in size. In other words, differences that surpass the WWC criteria are likely to occur simply by chance. Characteristics of study designs that enhance their scientific value, such as the use of multiple dependent measures, increase the probability that differences would surpass the criterion. Randomized control trials, the method most preferred by the WWC, typically have smaller samples. Yet such smaller samples also have increased probabilities of baseline differences that surpass this criterion. (See Appendix A for more details.)

Interestingly, the criterion does not include the possibility of comparing baseline differences with the effect of an intervention. In other words, even if posttest differences are substantially greater than the baseline differences the study cannot be considered. For example, Hattie (2009, pp. 206-207) found, in his summary of meta-analyses, that the average effect size associated with Direct Instruction reading programs was .89. This value is more than three times the cut-off of baseline differences stipulated by the WWC. Suppose that an intervention group began a study with average scores that were .30 of a standard deviation below a comparison group, a difference that would result in exclusion based on the WWC criteria. Hattie's results suggest that this intervention group, if given a DI reading program such as *RM*, could end the intervention with scores that were an average of .59 of a standard deviation above the comparison (.59 = -.30 + .89). Such a difference falls well within the range of results considered educationally significant, yet the study would be

rejected for consideration because of the pretest differences. (The companion technical report has examples of WWC decisions that resemble this scenario.)

Surprisingly, the article by Ho and associates (2007) that is cited by the WWC to support their criteria of group equivalence actually addresses methods to use matching techniques to enhance validity of results and does not appear to provide supporting evidence for their use of this standard. Ho, et al. briefly mention the .25 standard deviation criterion in one sentence in a footnote with a reference to Cochran (1968) (see footnote 15 on page 20 of Ho, et al, 2007). The 1968 Cochran piece also discusses matching procedures and a technique of adjustments of subgroups to develop equality, but provides no discussion or defense of the criterion used by the WWC. Most notably, the Cochran article was written before the development of computer technology that allows the relatively easy use of advanced statistical adjustments. Cochran notes that the "objective [of the adjustment process] is the same as in a standard analysis of covariance" (p. 301) and that "for samples of any reasonable size, covariance gives greater gains than stratified matching" (the technique he discusses in the article) (p. 309). This conclusion is also presented in Cochran and Chambers (1965). Thus, neither the work of Ho, et al (2007) nor that of Cochran, who is cited by Ho and associates, appears to provide the justification for the WWC criterion.

In addition, Ho and associates explicitly note that studies employing random assignment need not worry about further matching (pp. 205-206). Yet, as documented in Appendix A, the WWC has used the criterion regarding the magnitude of group differences to discard studies that used random assignment. Again, simple sampling theory tells us that group differences larger than a predetermined amount will appear by chance even with random sampling. It is hard to envision a logical reason for trying to improve on random assignment, a conclusion echoed by both Ho, et al. (2007) and Cochran (1968).

In short, simple statistical probabilities result in the WWC criterion regarding group equivalence being extraordinarily difficult to meet when studies use numerous measures or when they have smaller samples typical of randomized designs, both of which are encouraged by other WWC criteria. They also fail to consider the magnitude of a study's effects relative to the baseline; and a justification for this WWC criterion does not appear to be included in the work cited in their guidelines.

Non-Confounded Design – A prime concern of the methodological literature on research design is promoting internal validity, defined by Campbell and Stanley, as "the basic minimum without which any experiment is uninterpretable" (1963, p. 5). The WWC *Procedures and Standards Handbook* uses the term "confounds" to describe several issues related to internal validity:

> In quasi-experimental design studies, confounding is almost always a potential issue due to the selection of a sample because some unobserved

> factors may have contributed to the outcome. The WWC accounts for this issue by not allowing a quasi-experimental design studies (sic) to receive the highest evidence rating (WWC, 2014, p. 20)

Ironically, however, as noted above, randomized control designs in school settings can actually have more confounds than the alternative designs that the WWC does not accept.

The WWC's *Procedures and Standards Handbook* discusses a number of possible confounds to designs. One aspect involves criteria related to "one unit per condition":

> The most common type of confounding occurs when the intervention or comparison group contains a single study unit – for example, when all of the intervention students are taught by one teacher, all of the comparison classrooms are from one school, or all of the intervention group schools are from a single school district. In these situations, there is no way to distinguish between the effect of the intervention and that unit. For example, if all students who use a mathematics intervention are taught by a single teacher, then any subsequent differences between the outcomes of students who use the mathematics intervention and those who do not may be due to the intervention, the teacher, or both (WWC, 2014, p. 19).

In other words, if the design includes only one teacher, classroom, school, or district within a comparison group, it is considered a confounded design.

The example given regarding one teacher or one classroom within each unit is commonly accepted within the scholarly literature as a case of confounding, for it is logically impossible to separate the impact of an individual teacher from the intervention. However, the decision to exclude studies that involve comparisons between two separate schools or two districts, when there are data for multiple groups within these settings, is to our knowledge, quite unusual. Schools and districts typically have numerous classrooms and students, and thus many students taught by a variety of teachers. The scholarly literature includes many examples of well-regarded studies that have involved comparisons within a district or school. Advanced, multi-level, statistical analyses, similar to those employed in the next part of this paper, are also often used to control for such factors. Examples of using this criterion to exclude studies of *Reading Mastery* are discussed in the companion Technical Report (NIFDI Technical Report 2014-4).

The WWC also describes confounds that can occur from the nature of the implementation. One possibility involves potential combination of interventions:

> Confounding also occurs if an intervention is always offered in combination with a second intervention because any subsequent differences in outcomes cannot be attributed solely to either intervention. However, the WWC may view the combination as a single intervention and report on its effects (WWC, 2014, p. 20).

In other words, the protocol calls for the exclusion of studies in which the design involves the combination of interventions so that the unique contribution of a curriculum cannot be distinguished from other factors. This criterion is commonly accepted within the scholarly literature and is central to the CCSS discussions of internal validity.

It is important to note, however, that determining whether or not an intervention involves a confounding influence requires substantive knowledge of the intervention. An example is provided in the companion report of a WWC determination that key elements of the *RM* program, training of teachers and reinforcement of student responses, was actually a confounding factor and the reason for rejecting the study from consideration. Because teacher training and consistent student reinforcement are integral elements of the *RM* program it is reasonable to suggest that a reviewer with greater substantive knowledge of the subject would have avoided this error. Another example involved a review of a study of Reading Recovery (RR) that compared the standard RR program with RR combined with phonics instruction. The WWC chose not to use this comparison in their analysis claiming that the modification of Reading Recovery reflected an inappropriate confound (WWC, 2013c).[10]

A related issue involves the fidelity of implementation. If a treatment is not administered as designed, it is impossible to tell if an effect arises from the intervention itself or from a modification to the intervention. Surprisingly, however, the WWC criteria regarding the exclusion of studies for fidelity of implementation are relatively vague:

> Although the WWC documents how the intervention was implemented and the context in which it was implemented for the study sample, it makes no statistical adjustments or corrections for variations in implementation of the intervention (e.g., relative to an ideal or average implementation). Variations in implementation are to be expected in studies of interventions, since they take place in real-life settings, such as classrooms and schools, and not necessarily under tightly controlled conditions monitored by researchers….The topic area team leadership has discretion to determine whether these issues are substantive enough to warrant lowering the rating of a study or deeming it outside the scope of the review protocol. (WWC 2014, p. 21)

Other writings have analyzed the WWC's policies regarding the consideration of fidelity of implementation. Stockard (2010) describes logical flaws in the WWC's contention that variations in fidelity of implementation "balance out" and have no impact on assessments of the efficacy of an intervention results. Poor implementation of an ineffective program would result in assessments that provided overly positive reports, while poor implementation of an

[10] As would be expected, the study (Iversen & Tunmer, 1993) found that students made significantly faster progress when phonics instruction was included in the intervention.

effective program would result in assessments that provided overly negative reports. Thus, application of the WWC's approach would degrade good programs (false negatives) and inflate the efficacy of poor programs (false positives). While it is surprising that the WWC does not appear to have a standard policy or means of including fidelity of implementation as a criterion, the issue of fidelity was used in a determination regarding one study included in the November 2013 review of RM. (See Technical Report 2014-4.)

Finally, the WWC discussion of confounding factors includes a specific discussion of the comparison of data from multiple years:

> "…[I]f information on the treatment group comes from one school year, whereas information on the comparison group comes from a different school year, then time can be considered a confounding factor" (WWC, 2014, p. 20)

Ironically, use of this criterion explicitly excludes studies that use the "recurrent institutional cycle" or "cohort control group designs" that are recommended by the CCSS tradition for use in schools. The WWC discussion does not include a scholarly justification for this criterion. Again, the companion technical report has several examples of studies that appear to have been rejected on the basis of this criterion.[11]

Summary – Trying to Meet the WWC Criteria

In screening studies for possible inclusion in reviews the WWC appears to use a highly selective definition of appropriate designs, seemingly searching for a perfect experiment, akin to one that would be possible in a sterile laboratory environment with highly controlled conditions. This approach contrasts sharply with that traditionally used within the social sciences, exemplified by the writings of Campbell, Cook, Shadish and Stanley (CCSS). As described above, this standard methodological tradition is much more inclusive in nature, describing a variety of designs that are appropriate for complex organizational settings, such as schools; emphasizing the importance of both internal and external validity; and also stressing the importance of multiple tests of hypotheses using a variety of settings, samples, measures, and research designs. Moreover, the CCSS tradition emphasizes that there can be no "perfect" experiment and that a search for such perfection will generally not provide the most accurate information. Developing such a "perfect" design would be very difficult and expensive.

[11] This example of a confound was not included in earlier versions of the WWC's *Procedures and Standards Handbook*, nor in draft versions of the *Handbook* available through the end of 2013. Interestingly, the WWC statement does not appear to include the possibility of comparing changes over cohorts to changes within other groups. This variation of the cohort control group design, sometimes termed the "cohort control group with historical comparison design," controls for changes over time in a comparison group and thus could theoretically address the supposed confounding of time. In 2011 the NIFDI Office of Research and Evaluation transmitted to the WWC an article in the peer reviewed methodology journal *Quality and Quantity* that explicitly explained the logic of these designs and their suitability for educational research (Stockard, 2013, on-line first in 2011). The article includes extensive references to the CCSS tradition.

A simple thought experiment, in which one logically considers the various ramifications and possibilities, can illustrate the problems associated with designing a study of a curricular program that would meet all of the WWC criteria. To be accepted for review without reservations, one would need to use a pretest-posttest control group with randomized assignment design. To pass the "one unit" rule, one would need to ensure that neither the intervention group nor the control group had only one teacher, one classroom, one school, or one district. In other words, one would need to have a design that employed some type of multi-stage cluster sample. It appears that one would need at least four districts, randomly assigned to treatment and then one would potentially also need to randomly assign schools, classrooms, and students to avoid further violations of the one-unit rule. Note that one would also presumably need to only include districts that had more than one school and schools with more than one classroom at a grade. Thus, by definition, it appears that acceptable samples are limited to those that involve larger districts and schools, potentially eliminating many rural schools and smaller schools within urban areas. Then one would need to administer pretests to all of the students. Clearly, such a design would be enormously expensive and difficult to implement, especially if one wished to promote high fidelity of implementation of a program and to ensure that testing was conducted in a reliable manner.

While random assignment and pretesting could be theoretically accomplished with extensive funding, meeting the standard for group equivalence at pretest is subject to statistical probabilities and random error. Recall that the WWC standards require that the intervention and control group differ at pretest by less than .05 of a standard deviation on all measures, or by up to .25 of a standard deviation if acceptable statistical controls are used. This requirement is enforced even when studies incorporate random assignment. The WWC's protocol for studies of Beginning Reading lists four domains that should be assessed (alphabetics, fluency, comprehension, and general reading achievement). Two of these domains (alphabetics and comprehension) have sub-areas listed. Thus, a study that assessed all of the areas listed by the WWC would have measures of nine different skills. To meet the criteria for group equivalence at pretest one would need to ensure that pretest means of the experimental and control groups differed by less than .25 of a standard deviation on all nine measures (or .05 of a standard deviation if one did not use statistical controls).

As described above and discussed more fully in Appendix A, the probability of having group equivalence declines markedly when more variables are included in the analysis, but increases as sample size becomes larger. The number of cases needed to have a reasonable degree of confidence that one would meet the WWC requirements is not small. For instance, to be 90 percent confident that one would meet the more stringent .05 criterion on nine different measures, one would need, conservatively, a total sample of more

than 8,000 cases. If one wanted greater confidence, much larger samples would, of course, be needed.[12] (Appendix B explains the calculations used to derive these estimates.)

This small thought experiment illustrates why the CCSS tradition stresses that there can be no perfect experiment and why the scholarly methodological literature has, for decades, called for multiple tests of phenomena and the accumulation of results. Instead of devoting hundreds of thousands of dollars to trying to construct a "perfect" study, one could use the same amount of money to conduct dozens of studies in a wide variety of settings and with a range of outcome measures and designs. Such smaller studies would, of course, make it much easier to have greater control over the implementations and ensure true internal validity. Conducting studies in a range of settings would also promote external validity. Most important, having results from multiple studies would provide the accumulation of findings that the classical literature deems absolutely necessary for the advancement of science. The next section shows how such an accumulation could be used by the WWC. It also examines the utility of the criteria used by the WWC to judge the validity of studies.

Do The WWC Criteria Make a Difference? A Quantitative Analysis

As noted above, the November 2013 WWC report determined that none of the studies of *Reading Mastery* that they identified met their criteria for inclusion or standards for review. The companion document (NIFDI Technical Report 2014-4) provides details on all of the efficacy studies of *RM* that were rejected and documents reasons to question the WWC's decisions regarding 40 (53 percent) of the 75 efficacy studies that were examined.[13] The first sub-section below briefly summarizes the results of the studies in the WWC listing that could have been accepted for review. Mirroring other summaries of the literature on *Reading Mastery*, the vast majority of these works found strong evidence of its efficacy. The second sub-section looks at the relationship between the magnitude of reported efficacy and elements of the studies' designs. In other words, this analysis examines the relationship of the WWC criteria to conclusions about the program's efficacy. It asks, "Does the application of the exclusive approach to study reviews, like that used by the WWC, make a difference in conclusions regarding the impact of *Reading Mastery*?" The answer is, "No." Studies that met a given WWC criterion generally had very similar effect sizes as those that did not meet the criterion.

[12] The WWC also has strict guidelines regarding attrition, which have not been included in this brief thought experiment but which would place additional constraints on the probability that a single study would pass the criteria.

[13] The companion report also gives citations to 42 efficacy studies that appear to fit the review protocol but were not included in the WWC listing. All of these studies used a comparison group design recommended by the standard methodological literature, addressed the specified age and grade range, and were published in the time span considered by the WWC. Fifteen of these citations had been given to the WWC by NIFDI in previous communications.

Characteristics of Studies of *RM* that the WWC Could Have Considered

As described more fully in the companion report, each of the 40 studies for which the WWC determination could be questioned compared results for students who received *RM* to another group of students. All of the studies used a design recommended by the CCSS tradition. (Slightly more than half used some form of a post-test only control group design, while the others used a pretest-posttest design.) However, none of the studies met all of the WWC criteria. The most common reason for rejection was violating the "one unit" rule. Over four-fifths of the studies examined results from one district and over half examined results for groups within one school. Less than 10 percent had randomly selected samples, no doubt reflecting the difficulty of random assignment in field settings, a point noted, of course, by the CCSS tradition.

Effect sizes could be computed for 38 study designs within these articles.[14] Effect sizes are a common metric used to summarize the impact of educational interventions. They summarize the difference between an intervention and a control condition as a percentage of the common standard deviation (variability) of the two groups. An effect size of zero indicates no difference between the groups. Traditionally, effect sizes of .25 or larger have been seen as educationally important. Table 1 summarizes the distribution of average effect sizes found in the 38 analyses. They ranged from -.53 to 2.44, but only two of the 38 averages were less than zero. Twenty-eight (76%) were larger than the .25 criterion, and the average across all results was .57, more than twice the .25 level. This average value is similar in magnitude to the average effect sizes found in other meta-analyses involving *RM* and the DI approach as a whole (Coughlin, 2014; Hattie, 2009). (Appendix C gives citations to the studies and a summary of their design characteristics.)

Table 1
Effect Sizes in 38 Analyses of Reading Mastery with Beginning Readers Rejected for Inclusion in the WWC 2014 Report, Descriptive Statistics

Average	0.57
Minimum	-0.53
Maximum	2.44
Number with Effect < = -.25	1
Number with Effect > = .25	28

[14] Effect sizes were not calculated for a report using a single subject design or for one that used a norm comparison design, for at the time of writing this report the relevant norms had not yet been obtained. Both of these studies reported positive findings regarding the efficacy of RM and thus their inclusion would only strengthen the conclusions presented here. As explained more below and in the companion report, two reports incorporated two different designs and when design is used as the level two measure these are separated, resulting in a sample size of 38 separate designs. When "site" is used as the level two measure, these results are combined, as are others that have data from the same site. Results are consistent across the analyses.

The next section examines the extent to which variations in these effect sizes are related to the design characteristics of the studies. Does conformity to a given WWC criterion result in significantly different estimates of the effect associated with *Reading Mastery*? Do studies that come closer to a "perfect" experiment have significantly different estimates of the effect of *RM*?

Statistical Analysis of Variations in Effects

Mixed models were used to examine the way in which study characteristics were related to variations in effect size, modeling the analysis on work reported in Stockard (2013b). Mixed models are simply an extension of linear regression, but are especially useful when one has data on two or more levels—in this case multiple effect sizes in a given study.[15] They control for the multiple incidence of effects across studies by having the "study" variable as a random effect. Characteristics of each study can then be introduced as explanatory variables to determine the extent to which they impact estimates of the dependent measure, effect size.

There are two ways in which the level 2 categories can be defined, and results were calculated for both of these definitions. As explained in the companion report, one of the articles (O'Brien and Ware, 2002) was listed twice in the WWC document, perhaps because it analyzed the data in two different ways. Another (SRA/McGraw Hill, 2009) used two different data sets and designs. Some articles involved reports on the same school or district but were listed separately by the WWC. The first analysis below follows the WWC listing and treats each of the separate study designs as unique level 2 units. The second analysis uses the site from which the data were obtained, rather than the study report, as the level 2 distinguishing variable. Table 2 gives the descriptive data used for these two analyses; and it can be seen that the average values were very similar across the two groupings.

Tables 3 and 4 summarize the results of the mixed model analyses. Table 3 reports results when study design was entered as a random variable. The analysis included 38 designs, a total of 273 effects, one to 60 effect sizes reported in a design, and an average of 7.2 effects per study. The first line of data reports the results when there were no design characteristics in the model and the study design was entered as a random variable. The intercept value (the first column of data) reports the average effect size across the 38 studies adjusting for the differing number of reports from various studies. As would be expected, the value of .54 is very similar to the values given in Tables 1 and 2. More importantly, it is over twice as large as the value used to denote educationally important effects. In other words, averaging across these 38 designs the impact of *RM* on students'

[15] Mixed models are actually the regression equivalent of a nested factor analysis model.

reading skills is more than twice the level typically used to denote educational importance, a result that replicates other reports of the efficacy of *RM*.

Table 2
Descriptive Statistics, Variables Used in Meta Analysis

	Study as Random Variable		Site as Random Variable	
Variable	Mean	S.D.	Mean	S.D.
Pre-Post design	0.45	0.50	0.43	0.48
Random assignment	0.08	0.27	0.09	0.29
Statistical adjustments used	0.32	0.47	0.31	0.45
Cohort control group design	0.61	0.50	0.61	0.48
Other school used as control	0.16	0.37	0.15	0.36
Only one school	0.55	0.50	0.61	0.50
Only one district	0.84	0.37	0.86	0.34
Pretest diffs. > .25 s.d.	0.11	0.31	0.10	0.30
Rejected because of Design	0.63	0.49	0.66	0.46
Number of Students	5304	9229	-----	-----
Average effect size	0.57	0.50	0.61	0.51

Note: For the analyses with study as the random variable, there were 38 cases for each measure except number of students, where n=17. For the analysis with site as the random variable, there were 33 cases for all measures. Data on number of students were not available for the analysis with site as a random variable, because it varied from one study using a site to another. The average effect size varied, in both analyses, from -.53 to 2.44. The number of students ranged from 31 to 22,078. All other variables were dummy variables (coded 0,1).

This "intercept only" model, in the first line of Table 3, is used as the baseline against which all the subsequent models in the table are compared. Each of the following lines of data in Table 3 summarizes results of models that include one of the criteria used by the WWC to judge the validity of a research project. For instance, Model 2 (the second line of data) reports the impact of having a pretest-posttest design on the estimate of effect sizes. The value of the intercept (the first column of numbers) declines only very slightly (from .54 to .52) when this variable is entered and remains strongly significant. The second column gives the coefficient associated with the presence of a pretest. This value for Model 2 (0.04) is very small and not significant. (The value of .04 indicates that the estimated effect size is .04 higher for studies with a pretest than for those without a pretest.) The third column of data gives the -2 Log Likelihood (-2LL) values. These are a measure of model fit, which have a chi-square distribution and can be used to compare the relative fit of a given model to a less complex one. The fourth column of data has these comparisons, reporting the difference between the -2LL for each model and the baseline intercept model. For Model 2 the change is very small (0.04) and thus statistically insignificant.[16] In short, the results for

[16] Degrees of freedom for all comparisons are equal to one.

Model 2 indicate that the estimate of effect size remains essentially the same whether or not a study employs a pretest. Restricting analyses to only studies that include a pretest does not alter the nature of the results. Thus, there appears to be no value added from the use of this criterion.

Table 3

Results of Mixed Model Analysis of the Relationship of Design Characteristics to Effect Sizes with Design Entered as a Random Variable

	b - intercept	b - control	-2*LL	change
1) Intercept Only	.54***	-----	500.78	
2) Pre-Post Design	.52***	0.04	500.74	0.04
3) Random Assignment	.49***	.57*	496.84	3.94*
4) Statistical Adjustments	.52***	0.04	500.72	0.06
5) Cohort Control Group	.43***	0.21	499.14	1.64
6) Other School Control Group	.62***	-.42*	496.4	4.38*
7) One School	.36***	.38*	495.12	5.66*
8) One District	.29[a]	0.31	498.34	2.44
9) Pre Differences GT WWC Standards	.52***	0.16	500.32	0.46
10) Rejected for Design	.61***	-0.13	500.24	0.54
11) Number of Students	.59***	$-0.13*10^{-4}$	499.42	1.36

* = p < .05. ** = p < .10, *** = p < .001, a = p < .10

Similar results appear with each of the other WWC criteria. All of the adjusted effect sizes shown in the first column of data surpass the .25 criterion of educational importance, and all but one is significant at well beyond the .001 level. The exception occurs with the "one unit" criterion and cases where data were obtained from only one district. (The coefficient of .29 surpasses the .25 level, but is only significant at the .10 level.) Only three of the control variables were statistically significant. These significant results indicate that effects were somewhat higher for studies using random assignment and occurring within one school, but lower for those that had other schools as the control group. Only these three models (#3, 6, and 7) had significantly better fit than the base-line intercept only model. Importantly, however, the effect size associated with *RM* in each of these models remained well above the level seen as educationally important and is statistically significant.

Table 4 reports results of the mixed model analyses when the site was used as the random variable. For this analysis there were 33 sites, with 1 to 64 effect sizes calculated per site and an average of 8.3 effects.[17] Only two of the variables associated with the WWC criteria

[17] Within sties, some variables (the presence of a pre-post design, statistical adjustments for equivalence, use of a cohort-control group design, violations of the one-unit rule at the district level, and differences in pretest

had statistically significant influences on the effect sizes, using another school as a control group and having data from only one school, but the impact of these controls was in opposite directions. More important, all of the adjusted effect sizes were well beyond the level usually designated as educationally important and all were statistically significant.[18]

Table 4
Results of Mixed Model Analysis of the Relationship of Design Characteristics to Effect Sizes with Site Entered as a Random Variable

	b - intercept	b - control	-2*LL	change
Intercept Only	.57***	-----	497.32	
Pre-Post Design	.52***	0.10	496.68	0.64
Random Assignment	.51***	0.55	493.96	3.36
Statistical Adjustments	.53***	0.11	496.56	0.76
Cohort Control Group	.50***	0.13	496.52	0.8
Other School Control Group	.65***	-.44*	493.56	3.76[a]
One School	.38***	.37*	492.94	4.38*
One District	.34*	0.29	495.16	2.16
Pre Differences GT WWC Standards	.56***	0.05	497.24	0.08
Rejected for Design	.67***	-0.15	495.8	1.52
Number of Students (site average)	.61***	$-0.11*10^{-4}$	496.6	0.72
Number of Students (by design)	.67***	$-0.13*10^{-4}$	496.2	1.12

* = p < .05. ** = p < .10, *** = p < .001, a = p < .10

Taken together, these results suggest that criteria used by the WWC to exclude studies from consideration – those that are believed to characterize an exemplary study – actually have very little impact on estimates of the effect size associated with an intervention. The estimate of the effect of *Reading Mastery* on students' reading skills is similar no matter what type of design characteristic is considered – the presence of pretests, the use of random assignment or of statistical adjustments to enhance comparability, or having pretests that exceed the WWC defined limits. The results are also similar when studies do or do not violate the "one unit" rule, gathering data from one school or one district or when using another school as a control group. Finally, the results are similar no matter how many students are included in the analysis. The conclusion that *Reading Mastery* has an educationally important and statistically significant impact on students' reading achievement remains very robust under any of these considerations. The only exception reflects one study out of the 38 included that had a substantial negative effect. Such a negative effect would be anticipated simply by chance. These results replicate an earlier

scores that exceeded the WWC criterion) varied at the site level. Note also that number of students is measured in two ways, one as an average of values for the site and another as the value associated with each design. The latter measure is the one that could potentially vary by design within a site).

[18] As noted in the companion Technical Report (2014-4), one could argue that the reports numbered 8 and 20 in Appendix E should have been excluded from consideration at Step 2 using the WWC protocol. Results of the mixed models summarized in Tables 3 and 4 were identical when these two studies were omitted.

analysis (Stockard, 2013b) that examined 20 studies of *Reading Mastery* for students with learning disabilities.

Summary and Discussion

The educational research literature includes thousands of studies regarding the efficacy of intervention programs. Numerous meta-analyses and literature reviews have summarized this literature. The literature on Direct Instruction programs is especially large. For instance, in his meta-analysis of meta-analyses Hattie (2009) reported the results of 304 studies of Direct Instruction, involving over 42,000 students. Other reviewers have commented specifically on the size and quality of the evidence related to Direct Instruction programs, including *Reading Mastery* (e.g., Borman, et al., 2013). Despite this large literature base, the What Works Clearinghouse has identified very few studies that meet its selection criteria and evidence standards. For instance, the November 2013 review of *Reading Mastery* for Beginning Readers reported finding no studies published since 1983 that could meet their standards. Why, given such a large literature base, does the WWC accept so few studies for review? This technical report examined that question.

The first part of the report compared the WWC's procedures with those that are standard within the social sciences, contrasting the WWC's exclusive approach to the much more inclusive methodology commonly used in the social sciences and typified by the writings of Campbell and Stanley and their successors (Campbell and Stanley, 1963, Cook, Campbell, & Stanley, 1979; Shadish, Cook, and Campbell, 2002 and termed the CCSS tradition in our discussion above). For instance, the WWC accepts only studies that use a pretest-posttest control group design, giving the highest ranking only to those that include random assignment. In contrast the standard methodological tradition recommends a much wider range of designs, especially for organizational settings such as schools. This literature offers a number of alternatives to random assignment, noting that it is often not feasible and, in fact, can diminish internal validity of a study in organizations and groups in which participants routinely interact with each other. The standard CCSS tradition also downplays the need for pretesting, showing a number of viable, and often preferable, alternatives. In short, while the WWC has a very limited definition of acceptable research designs, the standard methodological literature provides a flexible and broad set of alternative approaches that can and should be used in field settings such as those in education.

Even if a study uses a pretest-posttest control group design, it must pass additional standards to be considered by the WWC, some of which appear quite difficult to meet. For instance, the WWC requires that pretest scores of an intervention and control condition be within .25 of a standard deviation of each other on all measures. The randomized trials preferred by the WWC are often small and the WWC requires, for full endorsement, that numerous measures be included. Yet, simply by chance, the probability of meeting the WWC standard for group equivalence declines as samples become smaller and as more measures

are included. Another problematic standard involves the "one unit" rule, which requires that studies include data from more than one classroom, school, or district, even if the classroom has multiple teachers, the school has multiple classrooms or data from multiple years, or the district has multiple schools. A large proportion of studies are automatically discarded based on this standard. Interestingly, the WWC provides no justification from the research literature for these standards or other requirements.[19]

The result of this process is that the WWC finds very few, and often no, studies that are deemed worthy of consideration for a given report. A thought experiment detailed above illustrates why it is so difficult to meet the WWC's standards by postulating characteristics of a "perfect" study that would include the preferred pretest-posttest control group with randomized assignment design; incorporate sufficient districts, schools, and students to pass the "one unit" standard; and have enough subjects to have a reasonable probability of meeting the requirements of group equivalence. Such a study could be extraordinarily expensive. Moreover, given statistical realities, it would have no guarantee of passing the various WWC criteria, especially that involving group equivalence. While each WWC criterion and standard might on its own appear logical or worthy, collectively they result in such severe restrictions that the probability of any study meeting all of these standards is very remote.

The standard methodological literature states quite firmly that there can never be a "perfect" experiment and instead calls for numerous tests of hypotheses, involving different settings, samples, and designs. This is the classic notion of a cumulative science. The second section of this report illustrates the use of such an accumulation of findings using 38 analyses of the use of the *Reading Mastery* curricular program with primary aged students, all of which were identified in the November 2013 WWC report. Results across these studies were very consistent. The average effect was more than twice the standard generally used to denote educational significance, a finding that replicates that commonly reported in the research literature on *RM*.

Multivariate statistical analyses were then used to examine the extent to which application of the WWC criteria affects substantive conclusions regarding the program's efficacy. The results indicated that there were no differences in the conclusions. There was no indication that effects are systematically smaller or larger when studies meet or do not meet a given criterion. The effects associated with *Reading Mastery* remained statistically significant and substantively strong when any of these criteria is used. There was no "value added" to the estimates of *RM*'s efficacy by considering the factors deemed important by the WWC, a

[19] As described earlier, the citation given by the WWC to justify the group equivalence criterion actually addressed a different issue and did not justify the criterion.

finding that replicated an earlier analysis of studies of the use of *RM* with students with learning disabilities (Stockard, 2013b).

On the other hand, the costs associated with the restrictions used by the WWC are far from minimal. As implied above, developing a study that would have a modest probability of passing the WWC standards could be extraordinarily expensive. In addition, the costs associated with the WWC screening process are no doubt very high with what appear to be a large number of person hours devoted to assessing studies for their eligibility. If the criteria were altered to more closely mirror those that are standard in the methodological literature, these resources could be reallocated to the review and summation of research findings as well as to more extensive checking of results.[20]

Just as important are the costs to the public. The use of the WWC's procedures deprives the public of a true view of the state of educational research. By suggesting that there are very few studies worth considering, they dismiss decades of solid research, much of it funded by the federal government and published in high ranking journals. Moreover, by devoting attention to only a few studies out of the many that are available, the WWC provides only a very small glimpse of the body of research results. Simple sampling theory would suggest that this very small sample could easily provide biased results. These biases will result no matter what curriculum is examined. Much more accurate estimates would, of course, be provided by a more complete sampling of the research literature.

Altering the WWC policies to align more closely with those of the traditional methodological literature would help to address this problem. Some of the most important needed changes include accepting the full range of research designs recommended for field settings, including schools; altering the standards regarding equivalence of groups at pretest; and applying the "one unit" rule only when it actually results in "confounded" results.
Most important, the WWC should always compare its conclusions with the extant scholarly literature, a regular step included in reviews that conform to standard methodological procedures. When differences occur it is incumbent upon the researcher to understand why they occur and to explain the discrepancies. However, as explained in correspondence with the NIFDI Office of Research, the WWC does not compare its results with those found by others.[21] Clearly, however, engaging in such comparisons could significantly increase the probability that the WWC reports would more accurately reflect the extant literature.

---

[20] The companion technical report documents a large number of errors at all stages of the review process in the November 2013 on RM, suggesting that much more attention needs to be paid to issues of accuracy in the review process. A forthcoming NIFDI technical report (2014-5) documents the extent to which numerous other groups and individuals have found such errors.

[21] The WWC procedures state that reviews should identify relevant existing systematic reviews and meta-analyses to ensure that we have identified all of the relevant literature. However, in response to a specific query regarding whether "the WWC reviews these meta-analyses while conducting their reviews and whether they consider including the differing opinions on the effectiveness of the programs reviewed in the reports,"

The WWC was funded to provide a cumulative picture of what works in education. But, the selection criteria that it employs embody such severe restrictions that it is virtually impossible to build a cumulative understanding of results. The standard methodological literature has well developed techniques and procedures for developing these cumulative understandings. The educational community, students and their parents, would be well served if the WWC would adopt these standard procedures.

the WWC responded (in an e-mail dated February 28, 2014), “The answer is no. As stated previously, other meta-analyses may differ in their inclusion criteria and standards. We do not report on or interpret the findings from other such reviews. We do list them in our citations so interested readers may find them." In other words, the WWC may review meta-analyses to identify studies, but, contrary to standard practices in the field, does not compare their results to those within the scholarly community.

Appendix A
The What Works Clearinghouse Criteria for Group Equivalence

As noted in the text, one of the common reasons that the WWC rejects studies for consideration is a conclusion that the intervention and comparison groups are not equivalent. This appendix uses sampling theory and the Central Limit Theorem to calculate the probability that studies could meet the WWC standards. The analysis shows how difficult it is for studies to meet the criteria and the ways in which having multiple measures or comparisons within a study and/or smaller samples increases the probability that a study will be excluded. Descriptions of two well-designed studies that were rejected for inclusion in recent WWC reports illustrate the nature of the problem. The appendix concludes with a brief discussion of implications.

Probabilities of Meeting the WWC Criteria

As described in the text, the WWC guidelines for determining if two groups that are compared in a study are equivalent require that information be provided on the groups' comparability prior to the intervention:

> If the reported difference of *any* (emphasis added) baseline characteristic is greater than .25 standard deviations in absolute value (based on the variation of that characteristic in the pooled sample), the intervention and comparison groups are judged to be not equivalent….For differences in baseline characteristics that are between .05 and .25 the analysis must include a *statistical adjustment* (emphasis in original) for the baseline characteristics to meet the baseline equivalence requirement. Differences of less than or equal to 0.05 require no statistical adjustment (WWC, 2014, p. 15).

The stipulations regarding the magnitude of acceptable baseline differences appear to be extraordinarily stringent. The paragraphs below use basic sampling theory to calculate the probability of researchers obtaining samples that would meet these criteria. The probability of meeting the criteria is not large and declines substantially when more than one measure is used and/or when sample sizes are smaller. Calculations are given for samples of size 64 and 36, typical of many studies examined by the WWC.

Example with Sample Size of 64 per Group

Suppose that a researcher were interested in selecting two random samples from a normally distributed population, with a mean of μ and a standard deviation of σ. Sampling theory tells us that if repeated random samples were drawn from this population, they would comprise a normal distribution, the sampling distribution, with a mean of μ and a standard deviation of $(\sigma/\sqrt{n})$, where n is the sample size. The standard deviation of the sampling distribution is called the standard error. In other words, the mean of an infinite number of drawn samples

equals the population mean, the standard error is a function of the standard deviation of the population and the sample size, and the distribution assumes the shape of a normal curve. We can use this logic (the Central Limit Theorem) to examine the probability that two randomly drawn samples, typical of those that the WWC prefers to have in studies that it examines, would have characteristics that met the criteria described above.

Consider a population with a mean of 50 and a standard deviation of 21, roughly equivalent to the Normal Curve Equivalent (NCE) distribution often used in education research. Suppose that a researcher drew two samples, each with 64 cases, from this population and designated one as the treatment group and one as the control group. For simplicity's sake we will assume that one of these samples (Sample A) perfectly represents the population, with a mean of 50 and a standard deviation of 21. (Note that this assumption is conservative in nature, resulting in the maximum probability of cases matching the WWC criteria.) To meet the WWC criterion of a difference at baseline of less than .05 of a standard deviation, the mean of the second sample (Sample B) would need to be within (.05 * 21 =) 1.05 points of the mean of Sample A, falling between 48.95 and 51.05. To meet the criterion of .25 of a standard deviation, Sample B would need to have a mean within (.25*21=) 5.25 points of the mean of Sample A, falling between 44.75 and 55.25.

We can use basic sampling theory to estimate how likely such an outcome would be. We begin by calculating the probability that one sample would be greater than .05 s.d. away from the population mean, assuming that the samples each have an n of 64. In other words, what is the probability that Sample B would have an average between 48.95 and 51.05, [P(48.95<=M<=51.05)]? Given that the mean of the sampling distribution would be 50 and the standard error (the standard deviation of the sampling distribution) would be $21/\sqrt{64} = 21/8 = 2.65$, we can calculate the z scores associated with each of these values:

$Z_{48.95} = (48.95 - 50)/2.65 = -.40$ and,

$Z_{51.05} = (51.05 - 50)/2.65 = +.40$.

Using a normal curve table we find that the probability of falling between these two values, or

P[48.95<=M<=51.05] = .1554 + .1554 = .3108.

Thus, the probability of choosing a sample that differs from the population mean (and by assumption the mean of Sample A) by less than .05 of a standard deviation is .31. The probability that the difference is larger than that amount, and thus subject to more stringent criteria by the WWC, is .69 (=1.00 - .31).

Suppose that the researcher was looking at three separate variables (e.g. fluency, comprehension, and vocabulary, similar to the dimensions reviewed by the WWC for analyses of beginning reading), each of which came from a population with an NCE distribution. Also assume that Sample A mirrors the characteristics of the population on each of these variables and that the sample size for each group (Sample A and Sample B) is

64. For each of the three separate measures the probability of obtaining a sample that fell within .05 s.d. of the population mean would be .31. But, having such an outcome for all three of the variables would only be .31*.31*.31 = .029. In other words, if a researcher were to examine three outcomes, the probability that all three of these scores would be within .05 of a standard deviation of the mean would be only .03. The probability that the measures would not meet the criteria would be .97. The probability of having samples that met the criteria when 5 measures are examined is much lower (p=.003). (See the first line of data in Table A-1.)

Table A-1

Probability that Samples Would Meet WWC Criteria Regarding Equivalence of Baseline Measures by Sample Size and Number of Dependent Variables

		Probability of Meeting Criteria		
Level	Sample Size	1 Measure	3 measures	5 measures
.05 s.d.	64	0.31	0.03	0.003
.25 s.d.	64	0.94	0.83	0.73
.05 s.d.	36	0.26	0.01	0.001
.25 s.d.	36	0.86	0.64	0.47

Note: As explained in the text, the probabilities are based on the assumption that one sample mirrors the population exactly. If this assumption did not hold, the probabilities of two samples meeting the criteria would be even smaller.

Consider the .25 of a standard deviation outcome, or the probability that the two samples would differ by more than .25 of a standard deviation and thus be rejected under the WWC criteria from any consideration, even with statistical adjustments. Again, using the NCE distribution, as described above, .25 of an s.d. = 5.25. So, paralleling the logic above, one may determine the probability that a sample would have an average between 44.75 and 55.25, or P[44.75<=M<=55.25]. Using the mean of the sampling distribution (50) and the standard error (2.65) given above, we can calculate the z scores associated with each of these values:

$Z_{44.75}$ = (44.75 - 50)/2.65 = -1.98 and,

$Z_{55.25}$ = (55.25 – 50)/2.65 = +1.98

Using a normal curve table we find that the probability of falling between these two values is .94, approximately equal to the standard 95 percent confidence interval. And thus, the probability of having a sample value that was greater than this level, given the sample size of .64, equals .06.

Again, however, with multiple dependent measures, the probability of falling outside the acceptable range becomes greater. Using the logic outlined above, if the probability of obtaining one sample that fell with .25 s.d. of the population mean is .94, having such an outcome for all three of the variables would be .94*.94*.94 = .83. In other words, if a

researcher were to examine three outcomes, the probability that all three of these scores would be within .25 of a standard deviation of the mean would be .83. There would be almost a 20 percent probability that at least one of the three measures would fail to meet the criteria and the study would then be excluded by the WWC. If the researcher examined 5 variables, the probability that the study would fall within the acceptable range (with all comparisons differing from Sample A by less than .25 s.d., but of course requiring additional statistical controls) would be .73. (See the second line of data in Table A-1.)

Example with Sample Size of 36 per Group

The issue becomes more difficult when samples are smaller, for then the standard error becomes larger. Let us assume that the sample size for each group is 36. Then the standard error = $21/\sqrt{36} = 21/6 = 3.5$, substantially larger than with the sample size of 64.

Consider first the criterion of having a sample mean within .05 s.d. (or +- 1.05) of the population mean. In this case,
$Z_{51.05} = 1.05/3.5 = .34$ and $Z_{48.95} = -1.05/3.5 = -.34$
Using a normal curve table one can find that there is a .26 (.13+.13) probability that the mean of Sample B, when the samples have an n of 36, will fall within .05 s.d. of the mean of Sample A, which, by definition, is equivalent to the population mean. The probability that the samples will differ by more than .05 s.d. is .74 (=1.00 - .26). If three measures are involved, the probability, with a sample size of 36 for each group, that all three measures will meet this criterion is only .02 (.0175). With more measures in the analysis the probability is, of course, even lower.

The last line of Table A-1 reports the same calculations for a sample size of 36 and utilizing the criterion of sample B differing by .25 s.d. from Sample A. With one measure the probability of meeting the criterion is .86, but the probability declines as more measures are included in the analysis. With 5 measures included the probability that a researcher using a sample size of 36 would meet the criterion of a difference of sample means of only .25 s.d. is less than .50.

In short, this demonstration shows how difficult it is for research projects to meet the WWC criteria regarding baseline equivalence of sample groups. The WWC strongly prefers randomized control trials, which are typically small, yet the probability of meeting the criteria declines as samples become smaller. In addition, the WWC reports include results from several measures. Yet, the more measures that a study reports, the greater is the chance that it will not meet the WWC criteria for baseline equivalence.

The Criteria in Practice

Two examples illustrate the way in which the WWC criteria have been applied. Both studies were well designed and published in top ranked journals in the field. Both were rejected for consideration by the WWC because the comparison groups were not equivalent at baseline.

The first is a quasi-experimental study, supported by grants from IES, the body that supports the WWC, and the National Institute for Child Health and Human Development (NICHHD) (Crowe, Connor, & Petscher, 2009). The study was rejected for consideration in the WWC's analysis of *Reading Mastery* for Beginning Readers because "it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent" (WWC 2013b, p. 2). Crowe and associates examined growth in reading achievement during one school year of over 30,000 students in grades one to three who were randomly selected from almost 3000 classrooms. They compared changes in oral reading fluency from fall to spring of students in six different curricula using growth curves and hierarchical linear modeling. The authors reported descriptive statistics at baseline on oral reading fluency for each of the groups in the analysis (p. 192) and for the total group. Of the 15 possible comparisons of a curriculum with *Reading Mastery (RM)*, the subject of the WWC report, three exceeded the .25 criterion set by the WWC. On average, the *RM* sample differed from the other groups by .12 of the total s.d., while the absolute value of the deviations ranged from .03 to .40. The fact that three of the fifteen comparisons exceeded the .25 s.d. level apparently resulted in the study being rejected, even though the statistical analyses nicely controlled for any baseline differences. (Interestingly, the pretest differences did not exceed the .25 criterion for one of the three grades examined, but the WWC rejected all of the results from consideration.)

The second example involves a randomized control design reviewed in the 2013 WWC analysis of *Reading Recovery*, a tutoring program for primary students at risk of future reading difficulties. The authors (Iversen & Tunmer, 1993) used a standard and well regarded method of controlled random assignment of subjects, matching 34 triplets of first grade students on beginning literacy skills and then randomly assigned members of the triplet to one of three groups. As with the Crowe et al. study, the WWC rejected this article for review because some of the differences between the intervention groups were greater than .25 of the pooled standard deviation (WWC 2013c, p. 34). The authors reported pretest data on 10 measures, resulting in 30 comparisons between the groups (Iversen & Tunmer, 1993, p. 119). Of the 30 comparisons between pretest means, 6 were larger than .25 of a standard deviation. The differences ranged from 0 to .55 of a standard deviation, and the average difference was .06 s.d. Using the methods described above, the probability that all 30 comparisons would be less than .25 s.d. is .005. The probability that all 30 comparisons would be less than .05 s.d., and thus not require further statistical adjustment, is very remote: 2.03^{-20}. In other words, simple calculations based on the Central Limit Theorem

would indicate a very small chance, given the sample size and number of comparisons, that Iversen and Tunmer's study would pass the WWC's criteria for baseline equivalence.

Summary and Discussion

Both of these examples illustrate how difficult it can be for a well-designed study to meet the criteria established by the WWC regarding equivalence of study groups. Even though the sample used by Crowe, et al. was extraordinarily large and used stringent and highly regarded analytic methods, differences emerged on a small proportion of the comparisons that were larger than the WWC set criterion. The chance of such differences emerging was, of course, heightened by the multiple comparisons included. The randomized design of Iversen and Tunmer had a substantially smaller sample than used by Crowe and associates (although, with a total n of 102, larger than that in many randomized studies) and employed a relatively large number of measures. The analysis in the first section of this paper shows how, simply by chance, differences that surpass the WWC criteria would be likely to occur. Ironically, while such multiple comparisons and careful designs make studies valuable within the academic research community, they greatly heighten the probability that they will not be accepted by the WWC.

The rational response of researchers who want to have their studies accepted by the WWC could be to limit their studies to very few measures and analyses. For instance, if Iversen and Tunmer had reported on only some of the measures in their analysis, their work would probably have met WWC acceptance criteria. If Crowe and associates had reported results from only one grade level, rather than three, their results would also have potentially been accepted. Yet, to have done so would have made their findings far less valuable for both educational researchers and the general public. It appears clear that, simply by statistical realities, the extraordinarily stringent requirements for group equivalence can result in the rejection of the vast majority of relevant studies and especially those that are more sophisticated.

A number of changes to the WWC criteria for group equivalence could address this issue. First, it would seem appropriate to accept all studies that use randomized assignment regardless of the extent of pretest differences. It is difficult to envision any rational, statistical justification for trying to improve upon randomization. Second, social scientists have developed sophisticated statistical techniques for analyzing data from quasi-experimental and field settings and adjusting for pretest differences, such as the growth models used by Crowe and associates. Given the power and wide acceptance of these approaches within the academic research community, it would seem reasonable to accept studies that use such statistical controls for analysis. In addition, because quasi-experimental, field based designs can have higher external validity than the more tightly controlled, yet often relatively artificial, characteristics of randomized control trials, the inclusion of studies with such statistical controls would potentially be of even greater

importance to the development of sound educational policy (McMillan, 2007, Stockard, 2013c). Third, given that the chance of finding cases that violate the criteria of group equivalence increases markedly with multiple measures and comparisons, it would be appropriate to look at the average difference across all measures and comparisons rather than omitting an entire study when only one difference surpasses a given level. Fourth, the WWC should consider any differences in pretest equivalence in conjunction with the magnitude of treatment effects. If treatment effects are large the criteria for similarity of baseline measures should be modified accordingly.

Appendix B
Calculations for Thought Experiment on Meeting Group Equivalence

This appendix reports the calculations used to determine the sample size needed to meet the WWC's criteria for group equivalence with varying numbers of dependent measures and level of confidence in the results.

Assume that we randomly assign students to two groups and that students can be randomly assigned across districts and schools, thus omitting the complications of multi-stage random sampling.[22] The calculations involved two steps.

The first is simply calculating the sample size (N) needed to assure groups fall within the needed levels of difference with defined probability levels. The sample size can be calculated using the formula for difference between the means, the classic t-ratio.

The standard error of the sampling distribution of the difference between two means is defined as
$\sigma_{M1-M2} = \sigma[\sqrt{(N_1+N_2)/(N_1*N_2)}]$, where σ is the standard deviation of the population and Ni is the size of the sample in group i.

If we assume $N_1 = N_2$, this reduces to
$\sigma_{M1-M2} = \sigma[\sqrt{(2N)/(N^2)}]$,

which can reduce further to
$\sigma_{M1-M2} = \sigma[\sqrt{(2)/(N)}]$. (1)

By definition, the t-ratio = $(M_1-M_2)/ \sigma_{M1-M2}$, where M_1 is the mean of sample 1 and M_2 is the mean of sample 2.

Suppose we are interested in drawing samples for which the difference between the means is .05 of the standard deviation, i.e. $M_1-M_2=.05*\sigma$. In that case, the t-ratio is equal to
$t = (.05*\sigma)/ \sigma_{M1-M2}$. (2)

Substituting from equation 1 above, equation 2 becomes
$t = (.05*\sigma)/[\sigma[\sqrt{(2)/(N)}]$.

The standard deviation values cancel out, and equation 2 can reduce to

[22] Note that this assumption is conservative in nature. Variability would, generally be greater in a sample selected in a multi-stage approach. Cluster samples almost always have more variation than simple random samples of the same size.

$t = [(.05*\sqrt{N})/\sqrt{2}]$.

Then, by simple algebra, one can derive the equation
$(\sqrt{2}*t)/.05 = \sqrt{N}$
And, with simple calculations
$28.284*t = \sqrt{N}$
And thus
$N = (28.284*t)^2 = 800 * t^2$ (3)

For the criterion of group differences of .25 s.d. or smaller, the required sample size can be calculated in a similar manner as
$t = (.25*\sigma)/[\sigma[\sqrt{(2)}/(N)]$.
The standard deviation values cancel out, and the formula can reduce to
$t = [(.25*\sqrt{N})/\sqrt{2}] = .1768 * \sqrt{N}$. (5)
By simple algebra, one can derive the equation
$\sqrt{N} = (\sqrt{2}*t)/.25 = 5.6569*t$
And thus
$N = (5.6569*t)^2 = 32 * t^2$ (6)

The formulas above give the sample size needed for a level of group equivalence when only one measure is involved. To calculate the sample sizes needed for multiple measures one can use the same logic, but focus not on the probability of one occurrence at a given level, but at the probability of multiple occurrences. Recall that the WWC requires that differences in pretest scores on all included measures must fall below .05σ (with no statistical adjustment) or .25σ (with statistical adjustment).

The probability of obtaining 2 results at a given level at the 90 percent confidence level would be .90*.90 = .81, the probability of 4 results at the 90 percent confidence level would be $.90^4 = .6561$. Based on this logic, if one wanted to be 90 percent confident that differences on 2 measures fell within a given range, one would need to actually have a sample size based on a $.90^{1/2}$ (=.948) level of confidence. With four measures, one would need to be calculate a sample size based on a $.90^{1/4}$ = (.974) level of confidence, etc. In other words, as explained in Appendix A, the more measures that are involved, the larger the sample must be if one is to meet the WWC criteria.

Table B-1 reports the t-ratios associated with varying levels of confidence and number of measures in the analysis. By definition, the t-ratio increases as the level of confidence increases (reading down the columns). Reading across the rows one can see the impact of having multiple measures. For instance, to be 90 percent confident that the differences between the mean meet a given criterion, one would need to be have a level of confidence

equal to .9740 if one considered 4 measures and a level of confidence of .9884 for 9 measures.

Table B-1

Levels of Confidence and Associated t-ratios by Number of Dependent Measures

One Measure		Four Measures		Nine Measures	
Confidence	t-ratio	Confidence	t-ratio	Confidence	t-ratio
0.900	1.65	0.9740	1.94	0.9884	2.27
0.950	1.96	0.9873	2.23	0.9943	2.53
0.980	2.33	0.9950	2.57	0.9978	2.84
0.990	2.58	0.9975	2.81	0.9989	3.06
0.999	3.29	0.9997	3.48	0.9999	3.69

Table B-2 reports estimates, derived using equations (3) and (6) given above, of the sample size needed to have differences that meet the WWC criteria of group equivalence at pretest assuming simple random assignment of cases and varying levels of confidence in the results. If one wanted to be 90 percent confident in the results, were only interested in one measure and planned to use statistical controls (the .25 criterion), one would need to have a total of 174 cases (87 in both the experimental and control group, see the first line in the second panel of Table B-2). However, if one wished to examine all nine dimensions one would need over 300 cases. To meet the .05 criterion one would need over 4000 cases to be 90 percent confident of meeting the criterion and close to 22,000 cases if one used all 9 measures and wished to be 99.9 percent confident.

Table B-2

Required Sample Sizes to Meet WWC Group Equivalence Criterion by Level of Confidence, Number of Dependent Measures, and Magnitude of Group Difference

Required Total Sample Size if M_1-M_2 < .05 s.d.			
Confidence	One Measure	Four Measures	Nine Measures
0.900	4,356	6,022	8,245
0.950	6,147	7,957	10,241
0.980	8,686	10,568	12,905
0.990	10,650	12,634	14,982
0.999	17,319	19,377	21,786
Needed Total Sample Size if M_1-M_2 < .25 s.d.			
Confidence	One Measure	Four Measures	Nine Measures
0.900	174	241	330
0.950	246	318	410
0.980	347	423	516
0.990	426	505	599
0.999	693	775	871

Note: The total sample sizes represent the sum of cases in the experimental and control group and assume simple random sampling from a population with a known standard deviation. Calculations for the .05 criterion are based on equation (3) and those for the .25 criterion are based on equation (6).

Appendix C
Studies Included in the Mixed Model Analysis

This appendix includes references to the studies included in the mixed model analysis and descriptions of their characteristics. All of the studies in the list were considered by the WWC in its 2013 report on *Reading Mastery* for Beginning Readers. Table C-1 summarizes the characteristics and results of these studies. The first column gives the study number, corresponding to the list of citations below. The next set of columns summarize characteristics of the studies that are related to the WWC criteria and standards including the presence of random assignment, use of multivariate statistical adjustments, a pretest-posttest control group design, the use of cohort control groups, characteristics related to the one unit rule (only two schools, only one school and only one district), and differences at pretest that exceeded the WWC limits. Also included is the number of students in the study, the step at which the WWC rejected the study and the number of effects that were calculated from the study. Full Descriptions of the studies are given in Appendices C and D of the companion technical report (2014-4)

1) Butler, P. A. (2003). Achievement outcomes in Baltimore City Schools. *Journal of Education for Students Placed at Risk, 8*, 33–60.

2) Fredrick, L. D., Keel, M. C., & Neel, J. H. (2002). Making the most of instructional time: Teaching reading at an accelerated rate to students at risk. *Journal of Direct Instruction, 2*(1), 57–63.

3) Gunn, B., Smolkowski, K., Biglan, A., & Black, C. (2002). Supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school: A follow-up. *Journal of Special Education, 36*(2), 69–79; and Gunn, B., Smolkowski, K., Biglan, A., Black, C., & Blair, J. (2005). Fostering the development of reading skill through supplemental instruction: Results for Hispanic and non-Hispanic students. *Journal of Special Education, 39*(2), 66–85.

4) Joseph, B. L. (2000). *Teacher expectations of low-SES preschool and elementary children: Implications of a research-validated instructional intervention for curriculum policy and school reform*. Dissertation Abstracts International, 65(01), 35A; and Vitale, M. & Joseph, B. (2008). Broadening the institutional value of Direct Instruction implemented in a low-SES elementary school: Implications for scale-up and school reform. *Journal of Direct Instruction, 8*(1), 1-18.

5) Marchand-Martella, N. E., Martella, R. C., Kolts, R. L., Mitchell, D., & Mitchell, C. (2006). Effects of a three-tier strategic model of intensifying instruction using a

research-based core reading program in grades K–3. *Journal of Direct Instruction, 6*(1), 49–72; and Marchand-Martella, N. E., Ruby, S. F., & Martella, R. C. (2007). Intensifying reading instruction for students within a three-tier model: Standard-protocol and problem solving approaches within a response-to-intervention (RTI) system. *TEACHING Exceptional Children Plus, 3*(5). We have not included this second analysis in our compilation of results, for the data are identical to those in the first paper.

6) O'Brien, D. M., & Ware, A. M. (2002). Implementing research-based reading programs in the Fort Worth independent school district. *Journal of Education for Students Placed at Risk, 7*(2), 167–195. (rejected by WWC for unacceptable design) (Design A)

7) SRA/McGraw-Hill. (2005b). *Barren County elementary schools post highest reading scores ever*. Columbus, OH: The McGraw-Hill Companies.

8) SRA/McGraw-Hill. (2005d). *Delaware charter school students maintain high reading scores*. Columbus, OH: The McGraw-Hill Companies.

9) SRA/McGraw-Hill. (2005e). *Direct Instruction helps Kentucky blue ribbon school attain record reading scores*. Columbus, OH: The McGraw-Hill Companies.

10) SRA/McGraw-Hill. (2005h). *Milwaukee elementary nearly doubles reading scores*. Columbus, OH: The McGraw-Hill Companies.

11) SRA/McGraw-Hill. (2005i). *Oregon Reading First project uses Reading Mastery Plus as core reading program*. Columbus, OH: The McGraw-Hill Companies.

12) SRA/McGraw-Hill. (2005j). *Phoenix inner-city students strive toward national reading average*. Columbus, OH: The McGraw-Hill Companies. (Note this was also listed by the WWC as SRA/McGraw Hill (n.d.m)

13) SRA/McGraw-Hill. (2005l). *Reading Mastery Plus helps Colorado school achieve AYP for first time*. Columbus, OH: The McGraw-Hill Companies.

14) SRA/McGraw-Hill. (2006a). *Cleveland school keeps Reading Mastery as curriculum core*. Columbus, OH: The McGraw-Hill Companies.

15) SRA/McGraw-Hill. (2006b). *DIBELS scores advance to grade level with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.

16) SRA/McGraw-Hill. (2006e). *Native American school uses Reading First grant to implement Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.

17) SRA/McGraw-Hill. (2007b). *Low-performing Kentucky school on its way to high-performing with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.

18) SRA/McGraw-Hill. (2007d). *Reading scores rise at Alabama elementary school with Reading Mastery Plus*. Columbus, OH: The McGraw-Hill Companies.

19) SRA/McGraw-Hill. (2007f). Title I schools in North Carolina district meet all-state reading targets with Direct Instruction. Columbus, OH: The McGraw-Hill Companies.

20) SRA/McGraw-Hill. (n.d.a) Seattle school boosts reading scores with Reading Mastery curriculum. Columbus, OH: The McGraw-Hill Companies.

21) SRA/McGraw-Hill. (n.d.b). *Anchorage school's diverse population flourishes with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.

22) SRA/McGraw-Hill. (n.d.n). "Nebraska District Outscores Peers Statewide, " pp. 14-15 in *Results with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.

23) SRA/McGraw-Hill. (n.d.p). *Success begins early at Alaskan elementary school*. Columbus, OH: The McGraw-Hill Companies.

24) Stockard, J. (2010). *The impact of Reading Mastery in kindergarten on reading achievement through the primary grades: A cohort control group design*. Eugene, OR: National Institute for Direct Instruction. (rejected by WWC for unacceptable design)

26) Brent, G., Diobilda, N., & Gavin, F. (1986). Camden Direct Instruction project 1984-1985. *Urban Education, 21*(2), 138–148.

27) Crowe, E. C., Connor, C. M., & Petscher, Y. (2009). Examining the core: Relations among reading curricula, poverty, and first through third grade reading achievement. *Journal of School Psychology, 47*, 187-214.

28) McIntyre, E., Rightmyer, E. C., & Petrosko, J. P. (2008). Scripted and non-scripted reading instructional models: Effects on the phonics and reading achievement of first-grade struggling readers. *Reading and Writing Quarterly, 24*(4), 377–407; and Rightmyer, E. C., McIntyre, E., & Petrosko, J. P. (2006). Instruction, development, and achievement of struggling primary grade readers. *Reading Research and Instruction, 45*, 209–241. These articles use the same design and data set.

29) O'Brien, D. M., & Ware, A. M. (2002). Implementing research-based reading programs in the Fort Worth Independent School District. *Journal of Education for Students Placed At Risk, 7*(2), 167-195. (Design B, rejected at step 3)

30) Stockard, J. (2011). Increasing reading skills in rural areas: An analysis of three school districts. *Journal of Research in Rural Education, 26*(8); and Stockard, J., & Engelmann, K. (2010). The development of early academic success: The impact of Direct Instruction's Reading Mastery. Journal of Behavior Assessment & Intervention in Children, 1(1), 2–24. Study B.

31) Carlson, C. D., & Francis, D. J. (2002). Increasing the reading achievement of at-risk children through Direct Instruction: Evaluation of the Rodeo Institute for Teacher Excellence (RITE). *Journal of Education for Students Placed At Risk, 7*(2), 141-166. Reprinted in *Journal of Direct Instruction, 3*(1), 29–50.

32) Jones, C. D. (2002). *Effects of Direct Instruction programs on the phonemic awareness abilities of kindergarten students*. Dissertation Abstracts International, 63(03), 902A.

33) Mac Iver, M. A., & Kemper, E. (2002). The impact of Direct Instruction on elementary students' reading achievement in an urban school district. *Journal of Education for Students Placed at Risk, 7*(2), 197–220.

34) Umbach, B., Darch, C., & Halpin, G. (1989). Teaching reading to low performing first graders in rural schools: A comparison of two instructional approaches. *Journal of Instructional Psychology, 16*(3), 112–121.

35) Ashworth, D. R. (1999). Effects of Direct Instruction and basal reading instruction programs on the reading achievement of second graders. *Reading Improvement, 35*(4), 150–156.

36) Green, A. K. (2010). *Comparing the efficacy of SRA Reading Mastery and guided reading on reading achievement in struggling readers*. Dissertation Abstracts International, 71(11A), 3969.

37) SRA/McGraw-Hill. (2009). *A report on the effects of SRA/McGraw-Hill's Reading Mastery, Signature Edition: A response to intervention solution*. DeSoto, TX: Author. (Note that there are two different designs in this study designated as 37a and b in Table C-1.)

38) Stockard, J., & Engelmann, K. (2010). The development of early academic success: The impact of Direct Instruction's Reading Mastery. *Journal of Behavior Assessment & Intervention in Children, 1*(1), 2–24. Study A.

Table C-1

Characteristics of Studies Using Design as the Level 2 Measure

Study Number	Average Effect Size	Random Assign.	Statistical Adjust.	Pretest-Posttest	Cohort Cont. Gp..	Two Schools	Only One School	Only One District	Not Equal at Pretest	Number of Students	Rejected at Step	Number of Effects
1	-0.11	No	No	No	No	Yes	No	Yes	No	4800	Two	60
2	0.58	No	Yes	Yes	No	No	Yes	Yes	No	107	Two	8
3	0.31	Yes	Yes	Yes	No	No	No	No	Yes	256	Two	47
4	0.60	No	No	No	Yes	No	Yes	Yes	No	1000	Two	18
5	0.13	No	No	Yes	No	No	Yes	No	No	184	Two	5
6	0.14	No	No	No	Yes	No	No	Yes	Yes	22078	Two	9
7	0.91	No	No	No	Yes	No	No	Yes	No	241	Two	5
8	0.79	No	No	No	Yes	No	Yes	Yes	No	40	Two	1
9	0.22	No	No	No	Yes	No	Yes	Yes	No	220	Two	2
10	0.56	No	No	No	Yes	No	Yes	Yes	No	146	Two	1
11	0.73	No	No	No	Yes	No	No	Yes	No	300	Two	2
12	0.85	No	No	No	Yes	No	Yes	Yes	No	320	Two	1
13	0.63	No	No	No	Yes	No	Yes	Yes	No	80	Two	1
14	0.79	No	No	No	Yes	No	Yes	Yes	No	90	Two	2
15	0.40	No	No	No	Yes	No	Yes	Yes	No	200	Two	3
16	0.90	No	No	No	Yes	No	Yes	Yes	No	136	Two	4
17	1.04	No	No	No	Yes	No	Yes	Yes	No	72	Two	1
18	0.11	No	No	No	Yes	No	Yes	Yes	No	131	Two	1
19	0.45	No	No	No	Yes	No	No	Yes	No	574	Two	2
20	0.61	No	No	No	Yes	No	Yes	Yes	No	96	Two	1
21	0.83	No	No	No	Yes	No	Yes	Yes	No	118	Two	1
22	0.50	No	No	No	Yes	No	No	Yes	No	1232	Two	4
23	0.48	No	No	No	Yes	No	Yes	Yes	No	164	Two	5
24	0.45	No	No	No	Yes	No	No	No	No	775	Two	2
26	0.97	No	Yes	Yes	No	No	No	Yes	No	119	Three	4
27	0.23	No	Yes	Yes	No	Yes	No	No	Yes	21003	Three	15

28	0.11	No	No	Yes	No	No	No	No	No	108	Three	13
29	0.26	No	Yes	Yes	No	No	No	Yes	No	22078	Three	12
30	0.57	No	Yes	Yes	Yes	No	No	No	No	1689	Three	6
31	0.79	No	Yes	Yes	No	Yes	No	Yes	No	20508	Three	17
32	0.49	Yes	Yes	Yes	No	No	Yes	Yes	No	36	Three	1
33	0.11	No	Yes	Yes	No	Yes	No	Yes	No	420	Three	4
34	2.44	Yes	Yes	Yes	No	No	Yes	Yes	Yes	31	Three	4
35	1.60	No	No	Yes	Yes	No	Yes	Yes	No	42	Three	1
36	-0.53	No	Yes	Yes	No	Yes	No	Yes	No	66	Three	2
37A	1.13	No	No	Yes	Yes	No	Yes	Yes	No	249	Three	4
38	0.44	No	Yes	Yes	No	Yes	No	Yes	No	169	Three	3
37B	0.23	No	No	Yes	No	No	Yes	Yes	No	33	Three	1

References

Adams G., & Engelmann, S. (1996). *Research on Direct Instruction: 25 years beyond DISTAR.* Seattle, WA: Educational Achievement Systems.

Agaodino, R. & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *The Review of Economics and Statistics, 86*, 180-194.

Biglan, A., Ary, D., & Wagenaar, A.C. (2000). The value of interrupted time-series experiments for community intervention research. *Prevention Science, 1*, 31-49.

Biglan, A. Flay, B.R., Komro, K.A., Wagenaar, A.C., Kjellstrand, J. (2012). *Adaptive time-series designs for evaluating complex multicomponent interventions in neighborhoods and communities*. Eugene, OR: Oregon Research Institute.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*(2), 125–230.

Brunswick, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.) Berkeley: University of California Press.

Campbell, D.T. (1986). Relabeling internal and external validity for applied social scientists. In W.M.K. Trochim (Ed.) *Advances in quasi-experimental design and analysis* (pp. 67-77). San Francisco: Jossey-Bass.

Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*:295-313.

Cochran, W.G. & Chambers, P.C. (1965). The Planning of Observational Studies of Human Populations. *Journal of the Royal Statistical Society. Series A (General)*, 128 (2), 234-266.

Cohen, B. P. (1989). *Developing sociological knowledge: Theory and method* (2nd ed.). Chicago: Nelson-Hall.

Cook, T.D., Scriven, M., Coryn, C.L.S., & Evergreen, S.D.H. (2009). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation*, published online December 4, 2009.

Cook, T.D. (1990). The generalization of causal connections: Multiple theories in search of clear practice. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (DHHS Publication No. PHS 90-3454, pp. 9-31). Rockville, MD: Department of Health and Human Services.

Cook, T.D. (1991). Clarifying the warrant for generalized causal inferences in quasi-experimentation. In M.W. McLaughlin & D.C. Phillips (Eds.), *Evaluation and education: At quarter-century* (pp. 115-144). Chicago: National Society for the Study of Education.

Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Coughlin, C. (2014). Outcomes of Engelmann's Direct Instruction: Research Syntheses, pp. 25-54 in J. Stockard (Ed.). *The Science and Success of Engelmann's Direct Instruction*. Eugene, OR: NIFDI Press.

Cronbach, L.J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Crowe, E. C., Connor, C. M., & Petscher, Y. (2009). Examining the core: Relations among reading curricula, poverty, and first through third grade reading achievement. *Journal of School Psychology, 47,* 187–214.

Engelmann, Z. (2014). Research from the inside: The Development and Testing of DI Programs, pp. 9-24, J.Stockard (ed.) *The Science and Success of Engelmann's Direct Instruction*. Eugene, OR: NIFDI Press.

Fisher, R.A. (1925) *Statistical Methods for Research Workers*. New York: McGraw-Hill.

Fisher, R.A. (1935). *The Design of Experiments*. London: Oliver & Boyd.

Glazerman, S., Levy, D.M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *Annals, AAPSS, 589,* 63-93.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London and New York: Routledge.

Heinsman, D.T. & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1,* 154-169.

Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15* (3), 199-236.

Iversen, S., & Tunmer, W. E. (1993). Phonological processing skills and the Reading Recovery program. *Journal of Educational Psychology, 85*(1), 112–126.

Kinder, D., Kubina, R., & Marchand-Martella, N. E. (2005). Special education and Direct Instruction: An effective combination. *Journal of Direct Instruction, 5*(1), 1–36.

Madigan, K. & Cross, R.W. (2012). *Impact of a schoolwide positive behavioral and intervention supports model on academic achievement in K-12 grades*. Accountability Works, Inc.

McMillan, J.H. (2007). Randomized field trials and internal validity: Not so fast my friend. *Practical Assessment, Research & Evaluation, 12* (15). http://pareonline.net/pdf/v12n15.pdf

O'Brien, D. M., & Ware, A. M. (2002). Implementing research-based reading programs in the Fort Worth independent school district. *Journal of Education for Students Placed at Risk, 7*(2), 167–195.

Popper, K. R. (1962). *Conjectures and refutations: The growth of scientific knowledge*. New York: Basic Books.

Przychodzin-Havis, A. M., Marchand-Martella, N. E., Martella, R. C., Miller, D. A., Warner, L., Leonard, B., et al. (2005). An analysis of *Corrective Reading* research. *Journal of Direct Instruction, 5*(1), 37–65.

Raudenbush, S.W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal, 45*, 206-230.

Schieffer, C., Marchand-Martella, N. E., Martella, R. C., Simonsen, F. L., & Waldron-Soler, K. M. (2002). An analysis of the *Reading Mastery* program: Effective components and research review. *Journal of Direct Instruction, 2*(2), 87–119.

Scriven, Michael. (n.d.). The Logic of Causal Investigations. Unpublished paper, Western Michigan University.

Scriven, M. (2008). A summative evaluation of RCT methodology: And an alternative approach to causal research. *Journal of MultiDisciplinary Evaluation, 5,* 11-24.

Shadish, W.R., Clark, M.H., & Steiner, P.M. (2008) Can randomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103,* 1334-1343.

Shadish, W.R.. Cook, T.D. & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Slavin, R.E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Research, 37*, 5-14.

SRA/McGraw-Hill. (2009). *A report on the effects of SRA/McGraw-Hill's Reading Mastery, Signature Edition: A response to intervention solution*. DeSoto, TX: Author.

Stockard, J. (2010). An Analysis of the Fidelity Implementation Policies of the What Works Clearinghouse. *Current Issues in Education*, 13 (4).

Stockard, J. (2013a). *Direct Instruction in the Guam Public Schools: An Analysis of Changes in Stanford Achievement Test Scores*, NIFDI Technical Report 2013-2. Eugene, OR: National Institute for Direct Instruction.

Stockard, J. (2013b) *Examining the What Works Clearinghouse and Its Reviews of Direct Instruction Programs*. Eugene, OR: National Institute for Direct Instruction.

Stockard, J. (2013c). Merging the accountability and scientific research requirements of the No Child Left Behind Act: Using cohort control groups. *Quality and Quantity: International Journal of Methodology, 47*, 2225-2257 (available on-line, December, 2011).

Tallmadge, G. K. (1977). *The Joint Dissemination Review Panel Ideabook.* Washington DC: National Institute of Education and U.S. Office of Education.

Tallmadge, G. K. (1982). An empirical assessment of norm-referenced evaluation methodology. *Journal of Educational Measurement, 19* (2), 97-112.

Weisburd, D., Lum, C.M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? *Annals, AAPSS,* 578, 50-70.

What Works Clearinghouse (2013a). About us. Washington, D.C.: Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/aboutus.aspx, retrieved September 9, 2013.

What Works Clearinghouse (2013b). *WWC intervention report, Reading Mastery and beginning reading*. Washington, D.C.: Institute of Education Sciences. Retrieved December 13, 2013, from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/WWC_ReadingMastery_081208.pdf

What Works Clearinghouse (2013c). *WWC intervention report, Reading Recovery and beginning reading*. Washington, D.C.: Institute of Education Sciences. Retrieved December 13, 2013, from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_readrecovery_071613.pdf

What Works Clearinghouse (2014). WWC procedures and standards handbook (Version 3.0). Washington, D.C.: Institute of Education Sciences. Retrieved August 13, 2014 from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

Zdep, S.M. & Irvine, S.H. (1970). A reverse Hawthorne effect in educational evaluation. *Journal of School Psychology, 8,* 89-95.